

Generalized Method for Probability-based Peptide and Protein Identification from Tandem Mass Spectrometry Data and Sequence Database Searching*

Antonio Ramos-Fernández‡, Alberto Paradela, Rosana Navajas, and Juan Pablo Albar§

Tandem mass spectrometry-based proteomics is currently in great demand of computational methods that facilitate the elimination of likely false positives in peptide and protein identification. In the last few years, a number of new peptide identification programs have been described, but scores or other significance measures reported by these programs cannot always be directly translated into an easy to interpret error rate measurement such as the false discovery rate. In this work we used generalized lambda distributions to model frequency distributions of database search scores computed by MASCOT, X!TANDEM with *k*-score plug-in, OMSSA, and InsPecT. From these distributions, we could successfully estimate *p* values and false discovery rates with high accuracy. From the set of peptide assignments reported by any of these engines, we also defined a generic protein scoring scheme that enabled accurate estimation of protein-level *p* values by simulation of random score distributions that was also found to yield good estimates of protein-level false discovery rate. The performance of these methods was evaluated by searching four freely available data sets ranging from 40,000 to 285,000 MS/MS spectra. *Molecular & Cellular Proteomics* 7:1748–1754, 2008.

Present day mass spectrometry-based proteomics research involves the generation of very large data sets containing thousands of tandem mass spectra, which are assigned to putative peptide sequences in databases by means of computer programs called database search engines. Given the number of MS/MS spectra involved, manual validation of spectrum to peptide assignments quickly became unfeasible, and user-unattended procedures for discarding incorrect matches were developed. In the earliest days of multidimensional chromatography coupled to tandem mass spectrometry, plain score cutoffs for each charge state were arbitrarily established by highly experienced mass spectrometrists (1, 2)

or determined by searching MS/MS spectra against reversed protein sequence databases (3). For instance, it was quite common to filter SEQUEST data by accepting all matches with $\Delta C_n \geq 0.1$ and $X_{corr} \geq 1.5$, ≥ 2 , and ≥ 3 for singly, doubly, and triply charged peptides, respectively. However, the relative frequency associated to a given score threshold was proven to be highly dependent on overall data set quality, database size, and database search parameters (4, 5). This finding implied that significance thresholds needed to be established in an experiment-specific manner and that score thresholds established for trial data sets should never be extrapolated to other data sets expecting that the error rate would be an experiment-independent variable uniquely associated to score values. Such concerns led to the development of mathematical models for describing the probability distributions of database search scores of commonly used search engines such as SEQUEST. Other researchers aimed at developing probability-based search engines attempting to directly provide a significance measure for each peptide assignment, such as X!TANDEM (6) or OMSSA¹ (7). Finally others decided to estimate error rates by comparing the frequencies of scores of peptide assignments with those obtained by assignments to false protein sequences obtained either by reversing or randomizing real protein sequences (8). Among these strategies, the recently described composite target/decoy sequence database search strategy is gaining increasing acceptance (9).

It is important to point out that warnings have been raised to encourage journals to increase the documentation of proteomics experiments, placing special emphasis on peptide and protein identification procedures, but current algorithmic diversity makes standardization a challenging task (for a detailed description of the current situation see a recent review by Nesvizhskii *et al.* (10)). Because of the number of database search engines available and the disparity of spectrum matching and scoring schemes that these programs implement, a

From the Proteomics Facility, Centro Nacional de Biotecnología, Consejo Superior de Investigaciones Científicas (CSIC), 28049 Madrid, Spain

Received, March 18, 2008

Published, MCP Papers in Press, May 31, 2008, DOI 10.1074/mcp.M800122-MCP200

¹ The abbreviations used are: OMSSA, Open Mass Spectrometry Search Algorithm; GLD, generalized lambda distribution; FDR, false discovery rate; DHR, decoy hit rate; iTRAQ, isobaric tags for relative and absolute quantification; InsPecT, interpretation of spectra with post-translational modifications.

generalized search engine-independent method to model score distributions and establish adequate statistical significance thresholds would be highly desirable. Furthermore statistical significance should be expressed in an easily interpretable form while allowing a good trade-off between stringency and power, such as the false discovery rate (FDR) (11, 12), which measures the expected proportion of truly null features that will pass a given p value threshold (i.e. the expected fraction of false positives).

In this work we used generalized lambda distributions (GLDs) to model MS/MS assignment score distributions. The GLD is an extremely flexible four-parameter function that can mimic with high accuracy the most important families of continuous probability distribution functions used in mathematical modeling. This distribution has been successfully used to model biological, physical, and economical processes because it can provide a valid mathematical model from observed data histograms of virtually any shape (13). By searching several collections of tandem mass spectra with a set of popular database search engines, either commercially or freely available, we demonstrate that this modeling strategy can be used in a search engine-independent manner to compute p values and peptide identification error rates. Finally we also provide a simple but powerful method for computing protein-level p values that are not biased for protein length or number of peptide hits. Estimates of associated protein-level identification error rates are also provided.

EXPERIMENTAL PROCEDURES

MS/MS Data Sets—All the data sets used in this work are freely available and contain MS/MS spectra recorded using ion trap mass spectrometers. The data set “RaftFlow,” containing approximately 40,000 dta files, was downloaded from the Sashimi documentation site (hosted by SourceForge). This data set corresponds to the analysis of the ICAT flow-through of lipid rafts purified from Jurkat T cells. The data set “PAe000038-39” was obtained by merging data sets PA000038 and PA300039 downloaded from the PeptideAtlas Web site that were obtained from proteome digests of human cancer cell lines SiHa and SqCC. MS/MS scans in mzXML files were converted to mgf file format as singly, doubly, and triply charged ions, yielding 53,666 spectra. The data set “PAe000114,” obtained from a digest of the human erythroleukemia K562 cell line, was also downloaded from PeptideAtlas. MS/MS scans in mzXML files were converted to mgf file format as singly, doubly, and triply charged ions, yielding 284,045 spectra. The data set “iPRG2008,” containing 42,235 MS/MS spectra, was obtained from the Association of Biomolecular Resource Facilities (ABRF) Proteome Informatics Research Group. These spectra were obtained from iTRAQ-labeled proteome digests of mouse liver cells.

MS/MS Database Searches—MS/MS database searches were carried out using MASCOT version 2.0.05 (available from Matrix Science under license), OMSSA 1.1.3.win32 (freely available from the National Center for Biotechnology Information (NCBI)), InsPecT “20070905” (freely available from the University of California Santa Cruz computational mass spectrometry group), and X!TANDEM 2 “2007.07.01.2” with k -score plug-in (freely available from LabKey). Peak lists in mgf format were used as input. Database search engine parameters were as similar as possible for all search engines. Briefly precursor mass tolerance was set to 2 Da, fragment ion mass tolerance was set to

0.8 Da, and cleavage specificity was set to “trypsin,” allowing for a maximum of two missed cleavages. Instrument-specific scoring was used when available. Cysteine alkylation due to iodoacetamide (+57.022) treatment was set as fixed modification for all data sets except for the iPRG2008 data set for which cysteines were treated with methylmethanethiosulfonate (+45.98). For this data set we also set as fixed modifications iTRAQ reagent (+144.102) at lysine side chain and peptide N terminus (except for InsPecT, which does not allow fixed modifications at the peptide N terminus). Databases used were target/decoy sequence databases built from International Protein Index (IPI) human v3.23 or the mouse UniProt database distributed along with the iPRG2008 data set. The use of random or reversed sequences as decoy proteins was decided arbitrarily.

Obtaining Database Search Engine Scores for Modeling—Matches to peptide sequences shorter than 10 residues were always discarded. For MASCOT, the main score was used; for X!TANDEM with k -score plug-in, the k -score (14) (an implementation of the COMET score (15)) was used; for InsPecT, we took the MQScore, which is a linear combination of several match quality scores computed by the program (16). Scores were used without modification but for OMSSA. In this case, to modify the scoring scale so that correctly identified peptides attain high scores, we took as score the negative logarithm of the reported “ E value.” GLD models were built for every charge state independently, and only assignments to reversed/random peptide sequences were used for this purpose. The number of data points was arbitrarily limited to the top 1500 scores of each charge state. This data set truncation was carried out to enforce the GLD model to minimize squared error in the right tail of the distribution for large data sets in the high score region where experimentally relevant statistical significance thresholds are usually computed. Fitting GLDs to truncated observed distributions is not a problem and does not affect model accuracy.

GLD Fitting—The generalized λ distribution is best defined from its percentile function,

$$Q(y) = Q(y, \lambda_1; \lambda_2; \lambda_3; \lambda_4) = \lambda_1 + \frac{y^{\lambda_3} - (1 - y)^{\lambda_4}}{\lambda_2} \quad (\text{Eq. 1})$$

where $0 \leq y \leq 1$. The parameters λ_1 and λ_2 are, respectively, the location and scale parameters, and λ_3 and λ_4 determine the skewness and kurtosis of the distribution. In the same way that the normal probability distribution has the restriction that the standard deviation must be non-0 and non-negative, not any given set of ($\lambda_1, \lambda_2, \lambda_3, \lambda_4$) parameters can yield a valid distribution. An accurate description of the restrictions on these parameters that yield a valid GLD may be found elsewhere (13). From the percentile function described above, the probability density at $x = Q(y)$ is computed as follows.

$$f(x) = \frac{\lambda_2}{\lambda_3 y^{\lambda_3} + \lambda_4 (1 - y)^{\lambda_4 - 1}} \quad (\text{Eq. 2})$$

Because y is defined as the probability of $x \leq Q(y)$, building data histograms for fitting a GLD involves turning data points into relative frequency scale, computing $Q(y)$ for all points, and binning data points according to this amount. For fitting GLDs to histograms from observed data we used the method of percentiles described by Karian and Dudewicz (13) with minor modifications. Briefly this method involves the computation of four sample statistics that are used as estimators of distribution parameters,

$$\hat{\rho}_1 = \pi_{0.5} \quad (\text{Eq. 3})$$

$$\hat{\rho}_2 = \hat{\pi}_{1-u} - \hat{\pi}_u \quad (\text{Eq. 4})$$

$$\hat{\rho}_3 = \frac{\hat{\pi}_{0.5} - \hat{\pi}_u}{\hat{\pi}_{1-u} - \hat{\pi}_{0.5}} \quad (\text{Eq. 5})$$

$$\hat{\rho}_4 = \frac{\hat{\pi}_{0.75} - \hat{\pi}_{0.25}}{\hat{\rho}_2} \quad (\text{Eq. 6})$$

where $\hat{\pi}_f$ denotes the sample percentile limiting the fraction f of ranked observations, and u is an arbitrary number between 0 and 0.25, which we set to 0.005. $\hat{\rho}_1$ is the sample median, $\hat{\rho}_2$ is the interdecile range, $\hat{\rho}_3$ is the left-right tail weight ratio, and $\hat{\rho}_4$ is the tail weight factor. Valid pairs of (λ_3, λ_4) parameters corresponding to the estimated amounts $(\hat{\rho}_3, \hat{\rho}_4)$, or $(1/\hat{\rho}_3, \hat{\rho}_4)$ if $\hat{\rho}_3 > 1$, are then obtained from previously tabulated values of (ρ_3, ρ_4) . From these amounts, the estimated values for λ_2 and λ_1 are obtained consecutively as follows.

$$\lambda_2 = \frac{(1-u)^{\lambda_3} - u^{\lambda_4} + (1-u)^{\lambda_4} - u^{\lambda_3}}{\hat{\rho}_2} \quad (\text{Eq. 7})$$

$$\lambda_1 = \hat{\rho}_1 - \frac{(1/2)^{\lambda_3} - (1/2)^{\lambda_4}}{\lambda_2} \quad (\text{Eq. 8})$$

Among all sets of $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ parameters compatible with the set of estimators $(\hat{\rho}_1, \hat{\rho}_2, \hat{\rho}_3, \hat{\rho}_4)$ obtained from each data histogram, the GLD $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ that better fits the observed data is selected so that it minimizes a given error indicator. An absolute error threshold of 0.35 in the sample estimators $(\hat{\rho}_3, \hat{\rho}_4)$ was used, and to bias the model toward better accuracy of the right tail, we selected as best model the one that minimized the amount

$$\sum_i w_i \frac{(y_i - f_i)^2}{f_i} \quad (\text{Eq. 9})$$

for every non-0 value of f_i where y_i and f_i denote, respectively, the value at the i th bin of the observed and model score histograms. The weight w_i was defined as the amount of 1 minus the relative rank raised to the power of 4. Optimal bin size for building histograms was computed from data set size, N , and interquartile range (a measure of data dispersion) as follows.

$$w = 2(\hat{\pi}_{0.75} - \hat{\pi}_{0.25})N^{-2/3} \quad (\text{Eq. 10})$$

Because a closed expression for the probability function of the form $y = F(x)$ does not exist, this distribution must be computed numerically to assign a p value to each data point.

Estimation of Error Rates in Peptide Identification—Given a set of p values assigned either at the peptide or protein level and ranked in ascending order, the expected proportion of data observations that pass a given p value threshold p_i depends on data set size and on the number of i data points with equal or better p value. This expected error rate is called FDR.

$$\text{FDR}_i = \frac{Np_i}{i} \quad (\text{Eq. 11})$$

Error rates may also be estimated from composite target/decoy sequence database searches by counting the number of decoy hits passing a given p value threshold. Because the FDR is a mathematical expectation, we decided to use a different name for the error rate estimated by counting decoy hits and called it decoy hit rate (DHR). This amount may be defined as

$$\text{DHR}_i = \frac{2D_i}{i} \quad (\text{Eq. 12})$$

where D_i is the number of decoy hits with p value better than p_i . Because this method is independent of the exact values of the estimated probabilities, it was used to assess the accuracy of the FDR computed from the models. Composite target/decoy databases were

built by concatenating a regular protein database with a fake version of it containing either random or reversed protein sequences. Random protein databases were built using the same distribution of protein lengths and residue frequencies observed in the original database.

Computation of Protein-level p Values and Error Rates—Peptide matches were grouped by parent protein sequence. From the p values of h peptide ions assigned to a given protein, the protein score was defined as

$$S_p = \sum_{i=1}^h -\log(p_i) \quad (\text{Eq. 13})$$

where p_i are the peptide ion p values computed from the corresponding GLD models. Because the null hypothesis states that the h_j peptides were assigned to the j th protein by chance, the probability of finding a score equal or better than S_{pj} may be estimated by generating a sufficiently large set of random protein scores and calculating the relative frequency of random scores achieving values as high as or higher than S_p . These random protein scores were computed by sampling h_j decoy peptide p values from the whole data set at random and applying Equation 13 to obtain a score. In this work we computed 10^6 random protein scores for each category of h in the data set, so protein p values as low as 10^{-6} could be assigned. Recall that lower p values may be assigned by increasing the number of random scores computed. After assigning p values, proteins were clustered into sequence similarity groups by defining a protein similarity cluster as the set of all proteins in the data set that shared at least one identified peptide. The FDR and DHR were calculated as described previously for each similarity cluster, taking as cluster p value the smallest protein p value in the cluster.

RESULTS

After database searches, MS/MS assignments to false peptide sequences were classified by charge state, redundancy was eliminated by taking the best scoring spectrum of a given charge state among those assigned to the same peptide sequence, and the resulting non-redundant set of scores was used to build a histogram of observed score frequencies. From every charge state-specific score histogram a model probability function was obtained by fitting a GLD as described under “Experimental Procedures.” These models were then used to compute the expected relative frequency of scores that will reach a magnitude as extreme as or more extreme than a given value, *i.e.* to assign p values. Fig. 1 shows GLD models obtained from database search scores provided by X!TANDEM with k -score plug-in, OMSSA, MASCOT, and InsPecT (from the data set RaftFlow). At first sight, probability density functions seem to provide a good description of score frequencies in all cases given the overlap between observed and model data series. To check the validity of the p values computed from the models, we pooled assignments from all charge states, ranked them by p value, and plotted the computed p value against its relative rank. As shown, both estimators converge to very similar values across several orders of magnitude, suggesting that the GLD fits may be used to estimate expected relative frequencies of raw

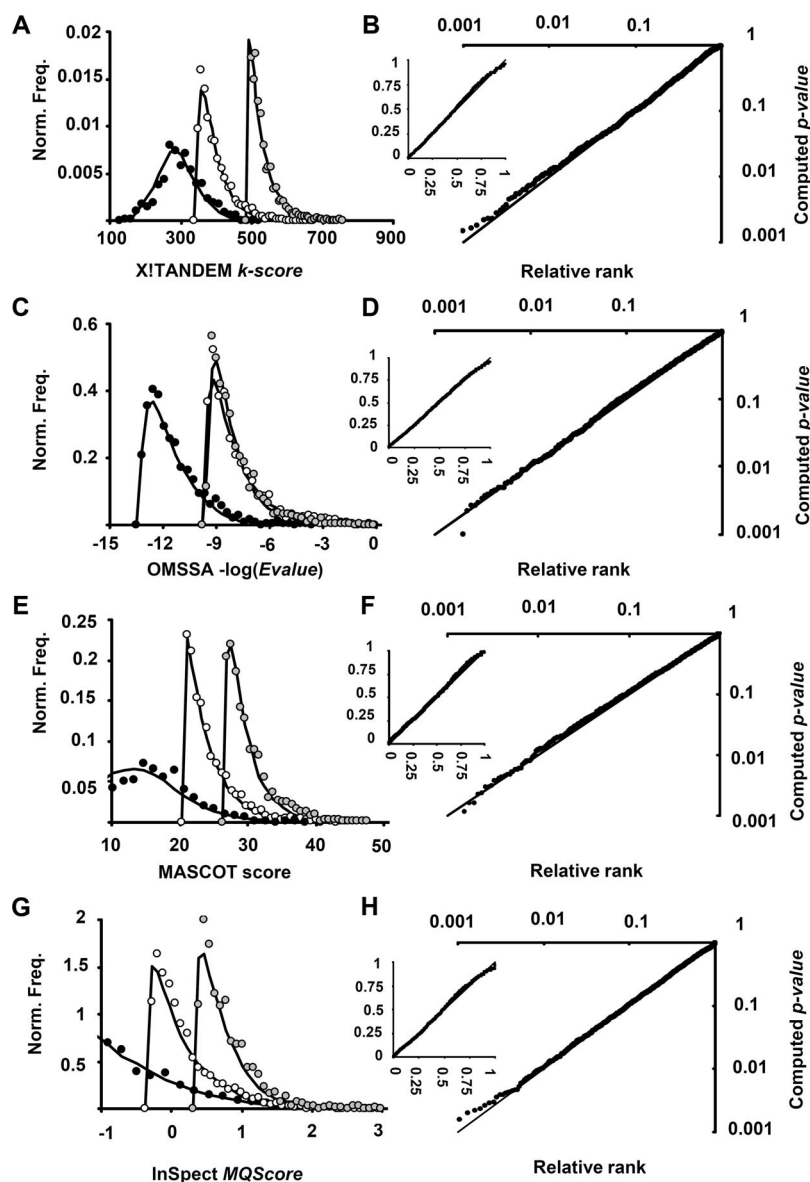


FIG. 1. GLD models of MS/MS assignments to decoy peptide sequences (pure random matches). A, C, E, and G show truncated score distributions for singly (black circles), doubly (open circles), and triply (gray circles) charged peptides with its best fit GLD model superimposed (solid lines) for X!TANDEM with *k*-score plug-in, OMSSA, MASCOT, and InSpecT, respectively. B, D, F, and H depict the comparison of *p* values computed from GLDs against the observed relative frequency of such *p* values (the solid line indicates the ideal relationship $y = x$) in logarithmic and decimal (insets) scales for the same set of search engines in the same order. Norm. Freq., normalized frequency.

scores with high accuracy. Furthermore we used the number of decoy hits passing a given *p* value threshold to estimate its associated frequency of random matches (DHR) and compared these amounts to FDR estimates obtained directly from the *p* values. The results obtained from searching the four data sets with the four search engines as well as parameters of the charge state-specific GLD models are shown in Tables I–IV. For a predicted fraction of incorrect matches of 5%, we observed frequencies of incorrect matches ranging from 3.7 to 5.5% in 15 of 16 cases and of 1.8% in one case. These results account for a total of 25,986 peptide ion matches. Afterward peptide matches were grouped by parent protein sequence, and a protein score was computed for every protein from the *p* values of putatively identified peptide ions. Given *h* peptide ions assigned to any one protein, by repeatedly sampling *h* decoy peptide *p* values at random and com-

puting a random protein score, we built protein score distributions that reflected the relative frequencies of the range of scores that may be obtained by combining *h* peptide ion *p* values under the null hypothesis. Such random protein score distributions are shown in Fig. 2A. As observed, score frequencies estimated from simulated distributions were very close to the observed frequencies of decoy protein matches in the data set, especially in the right tails. To check the validity of *p* values computed from these distributions, we pooled protein matches across all categories of *h* higher than 1, ranked them by ascending protein *p* value, and plotted the estimated protein *p* value against its relative rank. As observed in Fig. 2A, computed *p* values were very close to observed relative frequencies across several orders of magnitude. Because the absolute number of MS/MS hits at random to a given protein, *h*, depends on sequence length (larger proteins yield more theo-

Generalized Estimation of Misidentification Rates

TABLE I

Modeling results for the lipid raft ICAT flow-through data set, including parameters of the best fit GLD distributions for every charge state

“No. pept. ions,” number of peptide ions passing the arbitrarily specified FDR threshold (different charge states of the same peptide were counted separately) after combining peptides of all charge states. “Prot. FDR” and “Prot. DHR,” FDR threshold specified at the protein level and its corresponding DHR. “No. prot. clust.,” number of protein clusters passing the specified FDR threshold at the protein level.

Data set RaftFlow											
Engine/score	Charge	λ_1	λ_2	λ_3	λ_4	FDR	DHR	No. pept. ions	Prot. FDR	Prot. DHR	No. prot. clust.
MASCOT score	+1	11.48	0.0031	0.00579	0.0132	0.05	0.046	1406	0.05	0.033	401
	+2	19.49	0.0027	0.00039	0.0113						
	+3	25.56	0.0029	0.00039	0.0113						
InsPecT MQScore	+1	-1.59	0.1749	0.00431	0.15202	0.05	0.04	1832	0.05	0.033	416
	+2	-0.41	0.1597	0.00181	0.07966						
	+3	0.28	0.0269	0.00039	0.0113						
OMSSA -log(E value)	+1	-13.67	0.0722	0.00154	0.14887	0.05	0.042	1739	0.05	0.04	457
	+2	-9.95	0.0071	0.00039	0.0113						
	+3	-9.8	0.0077	0.00039	0.0113						
X!TANDEM k-score	+1	262.18	0.0044	0.15228	0.26478	0.05	0.044	1741	0.05	0.044	458
	+2	335.17	0.0014	0.00019	0.07795						
	+3	438.82	0.0003	0.00039	0.0113						

TABLE II

Modeling results for the combination of cancer cell line experiments PAe000038 and PAe000039

See legend in Table I.

Data set PAe000038-39											
Engine/score	Charge	λ_1	λ_2	λ_3	λ_4	FDR	DHR	No. pept. ions	Prot. FDR	Prot. DHR	No. prot. clust.
MASCOT score	+1	11.82	0.003	0.00039	0.0113	0.05	0.043	365	0.05	0.012	129
	+2	15.35	-0.0492	-0.0145	-0.127						
	+3	18.73	0.0031	0.00039	0.0113						
InsPecT MQScore	+1	0.607	0.6325	18.14	0.1711	0.05	0.018	354	0.05	0.025	121
	+2	0.033	0.0228	0.00039	0.0113						
	+3	0.324	0.1645	0.00181	0.07967						
OMSSA -log (E value)	+1	-13.67	0.1273	0.06768	0.3613	0.05	0.055	330	0.05	0.044	124
	+2	-9.73	-0.2191	-0.025	-0.1943						
	+3	-5.28	0.0081	0.00039	0.0113						
X!TANDEM k-score	+1	293.1	0.0095	12.1431	0.2879	0.05	0.049	236	0.05	0.059	101
	+2	260.9	0.0003	0.00013	0.0111						
	+3	440.9	-0.0059	-0.0106	-0.1247						

TABLE III

Modeling results for the large data set PAe000114

See legend in Table I.

Data set PAe000114											
Engine/score	Charge	λ_1	λ_2	λ_3	λ_4	FDR	DHR	No. pept. ions	Prot. FDR	Prot. DHR	No. prot. clust.
MASCOT score	+1	16.01	0.0031	0.00039	0.0113	0.05	0.04	3955	0.05	0.043	1196
	+2	20.13	0.0034	0.00039	0.0113						
	+3	20.38	0.0033	0.00039	0.0113						
InsPecT MQScore	+1	0.139	0.1676	0.00181	0.07966	0.05	0.037	4881	0.05	0.042	1138
	+2	0.759	0.0233	0.00039	0.0113						
	+3	1.051	0.1956	0.00353	0.08134						
OMSSA -log (E value)	+1	-10.42	0.0096	0.00039	0.0113	0.05	0.043	4433	0.05	0.043	1281
	+2	-8.75	-0.1091	-0.0106	-0.1247						
	+3	-8.26	0.0099	0.00039	0.0113						
X!TANDEM k-score	+1	248.9	0.0029	0.00019	0.07795	0.05	0.04	2579	0.05	0.043	908
	+2	310.3	-0.0049	-0.0145	-0.127						
	+3	457.7	-0.0062	-0.0106	-0.1247						

retical peptides upon *in silico* digestion) and a random score distribution was generated for every value of h , we expected that p values assigned by this method would not show any bias

toward parent protein length. This desirable behavior of protein p values was found to be true and is illustrated in Fig. 2B. Tables I–IV also show the results of protein-level postprocessing of all

TABLE IV

Modeling results for the iTRAQ-labeled mouse sample distributed by the Proteome Informatics Research Group

See legend in Table I.

Data set iPRG2008											
Engine/score	Charge	λ_1	λ_2	λ_3	λ_4	FDR	DHR	No. pept. ions	Prot. FDR	Prot. DHR	No. prot. clust.
MASCOT score	+2	20.57	0.0031	0.00067	0.01153	0.05	0.04	599	0.05	0.045	225
	+3	21.61	0.0035	0.00067	0.01153						
InsPecT MQScore	+2	−1.41	0.1219	0.00019	0.07795	0.05	0.052	297	0.05	0.099	134
	+3	−0.54	0.2779	0.00431	0.15202						
OMSSA $-\log(E \text{ value})$	+2	−13.2	0.0076	0.00739	0.01296	0.05	0.039	503	0.05	0.067	216
	+3	−12.94	0.0459	0.00181	0.07966						
X!TANDEM k -score	+2	235.25	0.0038	0.06351	0.18166	0.05	0.049	736	0.05	0.068	264
	+3	419.58	0.0004	0.00097	0.01176						

DISCUSSION

In previous research, other investigators developed statistical models to assign p values to peptide match scores provided by database search engines. These statistical models incorporate experiment-specific information that is reflected in the frequency distributions of such scores. In this work, we developed a generalized approach that relies on the use of generalized λ distributions, avoiding the need to search for a known probability density function that may approximate the frequency distribution of scores from a given search engine. To obtain models that minimize the squared error in the right tail, where meaningful significance thresholds are usually established, we truncated the score distributions by taking the 1500 highest scores of each charge state and introduced a weight factor in every normalized squared error that emphasized the contribution of the right tail to the total error. Because GLD models are extremely flexible, the complete absence of search engine scores coming from correctly assigned spectra is critical to prevent the model from also fitting the tail of positive match scores. This is the reason why we used only matches to false peptide sequences to estimate parameters for the models. Given the increasing popularity of the target/decoy composite database method, we preferred this strategy instead of separate real and false protein sequence database searches. In addition, this method allowed us to simultaneously obtain a widely accepted estimation of the misidentification rate to compare with that predicted from the statistical model. As described in under "Results," the accuracy of these models proved very good and yielded expected error rates very close to their real values in almost all cases. The fact that MS/MS search scores from any search engines may be expressed as p values using this method might facilitate a detailed comparative analysis of algorithm performance in the future. Moreover a common probability scale for all search engines seems much more suitable for the development of algorithms to obtain extra significance by matching MS/MS spectra to peptides using multiple database search programs.

Because identifying individual peptides is not the goal of most present day proteomics experiments, peptides must be

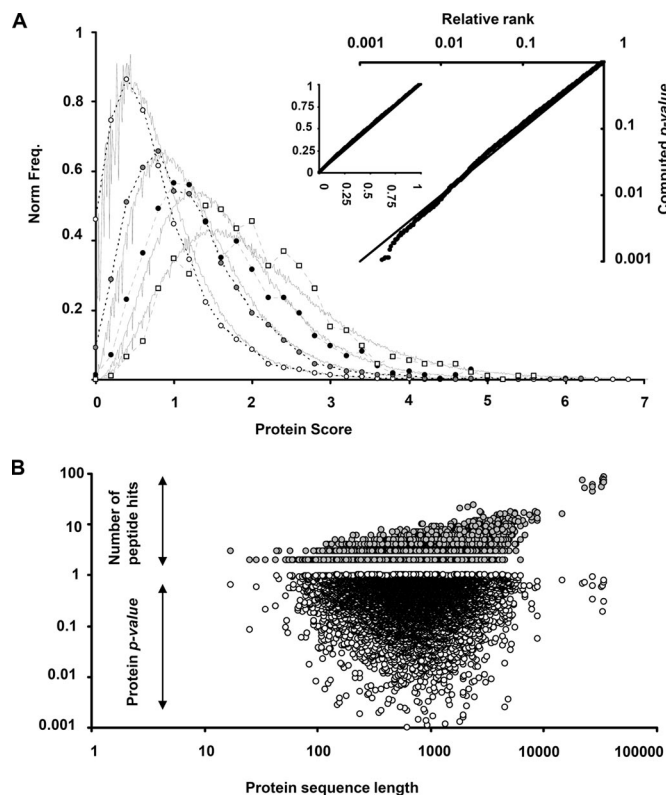


FIG. 2. A shows observed decoy protein score distributions for several values of h . ($h = 2$, open circles; $h = 3$, gray circles; $h = 4$, black circles; $h = 5$, open squares). Thin solid lines superimposed on observed distributions indicate distributions simulated by computing 10^6 random protein scores. The upper right subpanel shows the comparison of p values assigned from simulated distributions against the observed relative frequency of such p values (the solid line indicates the ideal relationship $y = x$) in logarithmic and decimal (inset) scales. B shows the number of peptide ion hits (gray circles) and protein p value (open circles) versus protein sequence length. Only data points belonging to decoy proteins (pure random matches) are shown. Norm. Freq., normalized frequency.

GLD modeling experiments described above. As observed, for a predicted 5% of incorrect protein matches, observed frequencies ranged from 3.3 to 6.7% in 14 of 16 cases and exceeded this range in only two cases with values of 1.2 and 9.9%. These results account for a total of 7569 protein clusters.

grouped according to parent protein sequence. For these groups of peptide matches, a new scoring scheme must be devised that represents each parent protein as a single experiment feature, and statistical significance thresholds must be established according to this new experimental level. To this end, we defined a protein score that reflected the quality of individual peptide ion matches by summing the negative logarithms of their p values. As described under "Experimental Procedures," this definition of protein score permits modeling the frequencies of random scores by chance in a straightforward manner and thus assigning protein-level p values without the need to develop an analytical formulation to estimate such probability. Estimation of protein p values and error rates using this method yielded values that were, in almost all cases, in very good agreement with the amounts estimated by counting decoy protein matches, suggesting that the method may be considered robust despite its simplicity. In addition, these protein p values are independent of the absolute number of MS/MS spectra matching peptides of a given protein, which is expected to be largely dependent on protein sequence length.

Recall that the ability to rank protein matches by decreasing significance is not intrinsic to the original target/decoy database strategy, which was initially defined at the peptide level. This added protein-level layer overcomes the fact that a misidentification rate at the peptide level may not necessarily be associated to a low misidentification rate at the protein level, which is a well known paradox of bottom-up proteomics. The other drawback of these proteomics technologies, known as the protein inference problem (17), was partially overcome by grouping proteins sharing identified identical peptide sequences into protein clusters. Because of sequence similarity among proteins, peptides released upon enzymatic digestion of a given protein may end up contributing to the putative identification of many other proteins not necessarily present in the sample. Although there is no perfect solution, a maximum parsimony assumption may be done to reduce the list of identified proteins to the minimum set of protein sequences capable of explaining the presence of all observed peptide matches. By taking the most significant protein match in each cluster as a representative match, protein clusters may be filtered under controlled error rate conditions as demonstrated in the tables. Attempting to do more sophisticated protein inference, for instance using non-degenerate peptides, is out of the scope of this work.

In conclusion, we think that the flexible statistical procedures presented might be applied to analyze MS/MS assignment scores from a variety of search engines as well as to obtain a non-redundant list of putatively identified proteins without significantly exceeding a maximum tolerated percentage of incorrectly identified proteins. The suitability of these procedures may be judged from the results obtained for the approximately 400,000 MS/MS spectra that were used in this work.

Acknowledgments—We are grateful to Alberto Medina for excellent informatics support and to Salvador Martínez de Bartolomé and Miguel Marcilla for critical revision of the manuscript.

* The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

‡ Recipient of an Itinerario Integrado de Inserción Profesional (I3P) fellowship from the CSIC (Ministerio de Educación y Ciencia).

§ To whom correspondence should be addressed: Centro Nacional de Biotecnología (CNB), C/Darwin 3, Universidad Autónoma de Madrid, 28049 Cantoblanco, Madrid, Spain. Tel.: 34-91-585-45-40; Fax: 34-91-585-45-06; E-mail: jpalbar@cnb.csic.es.

REFERENCES

1. Washburn, M. P., Wolters, D., and Yates, J. R., III (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **17**, 242–247
2. Florens, L., Washburn, M. P., Raine, J. D., Anthony, R. M., Grainger, M., Haynes, J. D., Moch, J. K., Muster, N., Sacchi, J. B., Tabb, D. L., Witney, A. A., Wolters, D., Wu, Y., Gardner, M. J., Holder, A. A., Sinden, R. E., Yates, J. R., and Carucci, D. J. (2002) A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* **419**, 520–526
3. Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J., and Gygi, S. P. (2003) Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J. Proteome Res.* **2**, 43–50
4. Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74**, 5383–5392
5. López-Ferrer, D., Martínez-Bartolomé, S., Villar, M., Campillos, M., Martín-Maroto, F., and Vázquez, J. (2004) Statistical model for large-scale peptide identification in databases from tandem mass spectra using SEQUEST. *Anal. Chem.* **76**, 6853–6860
6. Craig, R., and Beavis, R. C. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **9**, 1466–1467
7. Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., Yang, X., Shi, W., and Bryant, S. H. (2004) Open mass spectrometry search algorithm. *J. Proteome Res.* **5**, 958–964
8. Qian, W. J., Liu, T., Monroe, M. E., Strittmatter, E. F., Jacobs, J. M., Kangas, L. J., Petritis, K., Camp, D. G., II, and Smith, R. D. (2005) Probability-based evaluation of peptide and protein identifications from tandem mass spectrometry and SEQUEST analysis: the human proteome. *J. Proteome Res.* **1**, 53–62
9. Elias, J. E., and Gygi, S. P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **3**, 207–214
10. Nesvizhskii, A. I., Vitek, O., and Aebersold, R. (2007) Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* **10**, 787–797
11. Benjamini, Y., and Hochberg, Y. (1995) Controlling the false discovery rate: a practical powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300
12. Storey, J. D., and Tibshirani, R. (2003) Statistical significance for genome-wide studies. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 9440–9445
13. Karian Z. A., and Dudewicz, E. J. (2000) *Fitting Statistical Distributions: the Generalized Lambda Distribution and Generalized Bootstrap Methods*, Chapman and Hall/CRC, Boca Raton, FL
14. MacLean, B., Eng, J. K., Beavis, R. C., and McIntosh, M. (2006) General framework for developing and evaluating database scoring algorithms using the TANDEM search engine. *Bioinformatics* **22**, 2830–2832
15. Keller, A., Eng, J., Zhang, N., Li, X. J., and Aebersold, R. (2005) A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol. Syst. Biol.* **1**, 2005.0017
16. Tanner, S., Shu, H., Frank, A., Wang, L. C., Zandi, E., Mumby, M., Pevzner, P. A., and Bafna, V. (2005) InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal. Chem.* **77**, 4626–4639
17. Nesvizhskii, A. I., and Aebersold, R. (2005) Interpretation of shotgun proteomic data: the protein inference problem. *Mol. Cell. Proteomics* **4**, 1419–1440